CLUSTER SAMPLING WITH APPLICATIONS OF STIRLING NUMBERS OF THE SECOND KIND

by

Arthur J. Roth* and Milton Sobel*

Technical Report No. 217

University of Minnesota

* This research was supported by the National Science Foundation under Grant GP 28922X at the University of Minnesota.

1. THE PROBLEM:

For a linear chain (and also for a circular chain) consisting of n points, we consider various aspects of selecting c clusters of common size s (a cluster being a set of s consecutive points) from among the points in the chain. We obviously must require $s \geq 1$, $c \geq 1$, and $n \geq s$ for the problem to make sense. We assume an equi-probable distribution for selecting clusters, and we consider selecting both with replacement (allowing the same cluster to be selected more than once) and without replacement, in which case we clearly must have $c \leq n - s + 1$ for the linear chain and $c \leq n$ for the circular chain. In particular, we obtain the probability distribution of the number of points in the union (and also, separately, in the intersection) of the clusters selected, and for the intersection we also exhibit explicit formulas for the mean and the variance in all cases. It turns out that parts of the problem are related both to tail probabilities of the multinomial distribution and to Stirling numbers of the second kind; we use this relationship to introduce a new set of polynomials which generate these Stirling numbers. We remark that there is a continuous analogue to our problem which is also of some interest and which we intend to treat separately in the future, and we note that our present results can neither be derived from nor be used to derive the corresponding formulas for this analogue.

2. PROBABILITY THAT THE UNION K HAS EXACTLY k POINTS.

Before considering each case separately, we note that the result is zero in all cases if $k < s$. Hence, we consider only $s \le k \le n$.

For notation purposes, we let $P(k|n, c, s)$ mean $P\{K = k|n, c, s\}$ where the order of the given quantities n, c, s is important and each position is occupied by the symbol which had the meaning indicated in the introduction.

A. Linear Chain, With Replacement.

If $c = 1$, then $K = s$ with probability one. More generally it is clear that $P(s|n, c, s) = 1/(n - s + 1)^{c-1}$ for any c. So we assume that $c \ge 2$ and $s + 1 \le k \le n$, and we first consider separately the case $k = n$ (i.e., complete coverage of the chain). If we associate each cluster with its right endpoint, there are $n - s + 1$ possible points to choose from, and we label them $1, 2, \ldots, n - s + 1$. Then complete coverage is equivalent to the following two conditions:

(i) Points 1 and $n - s + 1$ are chosen at least once.

and (ii) There is no succession of s consecutive unchosen points (so that the difference between any two adjacent chosen points is at most s).

Since the $n - s + 1$ points have $n - s$ spaces between them, and since two of the chosen clusters are determined by (i), condition (ii) asserts that complete coverage is equivalent to partitioning the $n - s$ spaces into $c - 1$ non-negative integer parts (by selecting $c - 2$ additional clusters); a zero part corresponds to repeating a cluster. Let $\Psi(n - s, c - 1, s)$ denote the number of ordered ways we can select the c clusters to obtain a partition with properties (i) and (ii). Then it follows that

$$(2.1)\qquad P(n|n,\ c,\ s) = P(\text{complete coverage}) = \frac{1}{(n-s+1)^c}\ \Psi(n-s,\ c-1,\ s)$$

To obtain an explicit formula for the Ψ-function, let $A_0(n-s,\ c-1,\ s)$ denote the number of ordered partitions of $n - s$ into $c - 1$ parts, each part $\leq s$, with no zeros, let $A_1(n-s,\ c-1,\ s)$ denote the same with exactly one zero, let $A_2(n-s,\ c-1,\ s)$ denote the same with two adjacent zeros and no other zeros, let $A_{1,1}(n-s,\ c-1,\ s)$ denote the same with two non-adjacent zeros and no other zeros, etc. Then

$$(2.2)\qquad \Psi(n-s,c-1,s) = c!\left(A_0(n-s,c-1,s) + \frac{A_1(n-s,c-1,s)}{2!} + \frac{A_2(n-s,c-1,s)}{3!} + \frac{A_{1,1}(n-s,c-1,s)}{(2!)^2} + \ldots\right)$$

$$= c!\left(A_0(n-s,c-1,s)\{1\} + A_0(n-s,c-2,s)\left\{\frac{\binom{c-1}{1}}{2!}\right\} + A_0(n-s,c-3,s)\left\{\frac{\binom{c-2}{1}}{3!} + \frac{\binom{c-2}{2}}{(2!)^2}\right\} + \ldots\right).$$

To explain the last expression in (2.2) consider, for example, the term $A_{1,1}(n-s,c-1,s)$ in the middle expression. Since there are two non-adjacent zeros we consider the number of different positions that these zeros can occupy. For each ordered partition in $A_0(n-s,c-3,s)$ we can put the set of two zeros in $\binom{c-2}{2}$ possible positions, and hence the product of $\binom{c-2}{2}$ and $A_0(n-s,c-3,s)$ is the same as $A_{1,1}(n-s,c-1,s)$. A similar argument holds for each of the terms in the middle part of (2.2). Note that the braces in (2.2) multiplied by $c!$ yield integers that depend only on c, and hence have some interest per se. Also, the expressions in braces are polynomials in c; each is one degree higher than the previous one. We will denote by $P_i(c)$ the polynomial of degree i obtained in this manner for $i = 0, 1, 2,\ldots$ This notation will be used in Section 5 to show that these polynomials generate the Stirling numbers of the second kind.

It is well known that [1]

$$(2.3) \qquad A_0(t, p, m) = \sum_{i=0}^{[\frac{t-p}{m}]} (-1)^i \binom{p}{i} \binom{t-1-mi}{p-1}$$

where [x] denotes the integer part of x. Thus (2.1), (2.2), and (2.3) together yield an explicit formula for $P(n|n, c, s)$.

It remains to consider $s + 1 \le k \le n - 1$. Let $k' = n - k$, the number of points not covered by our c clusters, so that $1 \le k' \le n - s - 1$. We divide k' into p positive parts and determine how many ways we can insert these parts among the k points which are covered by the c clusters. Since dividing the k' points into p positive parts is equivalent to selecting $p - 1$ different spaces between these points, this can be done in

$$(2.4) \qquad D(k', p) = \binom{k' - 1}{p - 1}$$

ways; each way corresponds to different ordered partition of k'.

Let us assume we have a chain of size k rather than n, and we have complete coverage of this chain. In how many places can we insert one or more of the k' uncovered points without changing the (covered or uncovered) status of any of the points? We answer this question by considering the underlying reason for condition (ii) above, one of the necessary conditions for complete coverage. It becomes apparent that we can "break apart" the chain at one location for each pair of adjacent selected cluster endpoints whose distance apart is exactly s (rather than merely at most s), and that these are the only possible locations for our uncovered points besides the two ends of the completely covered chain of size k. Let $\Psi^{(j)}_{(n-s,c-1,s)}$ be the same as

$\Psi(n-s,c-1,s)$ except that exactly j of the parts are of size s. Note that we must have $j \le [\frac{n-s}{s}]$. Since the j parts of size s can be put in $\binom{c-1}{j}$ positions, and since the remaining parts can be chosen in only $(c-j)!$ different orders (instead of the original $c!$), we obtain for $j \le [\frac{n-s}{s}]$,

$$(2.5) \qquad \Psi^{(j)}(n-s,\ c-1,\ s) = \binom{c-1}{j} \frac{c!}{(c-j)!} \Psi(n-(j+1)s,\ c-1-j,\ s-1).$$

The result is, of course, zero for $j \ge [\frac{n}{s}]$. An explicit form for the Ψ-function in (2.5) can be obtained from (2.2) and (2.3) and leads to an explicit form for the $\Psi^{(j)}$-function.

The p positive parts of k' can each be inserted into any of these j places or at either end of the chain. If we use both ends, then we must choose $p-2$ additional locations out of the other j available ones. This can be done in $\binom{j}{p-2}$ ways. Similarly, if we use just one end of the chain the appropriate factor is $2\binom{j}{p-1}$, and if we don't use either end it is $\binom{j}{p}$. We must require $j \ge p-2$, $j \ge p-1$, and $j \ge p$, respectively, for the preceding three cases. Since the $\Psi^{(j)}$-functions represent disjoint events, we can add probabilities for each j to obtain for $s+1 \le k \le n-1$,

$$(2.6) \qquad P(k|n,c,s) = \sum_{p=2}^{k'} \sum_{j=p-2}^{[\frac{k-s}{s}]} \binom{k'-1}{p-1} \binom{j}{p-2} \frac{\Psi^{(j)}(k-s,c-1,s)}{(n-s+1)^c}$$

$$+ 2 \sum_{p=1}^{k'} \sum_{j=p-1}^{[\frac{k-s}{s}]} \binom{k'-1}{p-1}\binom{j}{p-1} \frac{\Psi^{(j)}(k-s,c-1,s)}{(n-s+1)^c}$$

$$+ \sum_{p=1}^{k'} \sum_{j=p}^{[\frac{k-s}{s}]} \binom{k'-1}{p-1} \binom{j}{p} \frac{\Psi^{(j)}(k-s,c-1,s)}{(n-s+1)^c}.$$

Using (2.5) and the fact that $k' = n - k$, we write finally for $s + 1 \le k \le n - 1$,

(2.7)

$$P(k|n,c,s) = \frac{1}{(n-s+1)^c} \sum_{\alpha=0}^{2} \sum_{p=\alpha}^{n-k} \sum_{j=p-\alpha}^{[\frac{k-s}{s}]} \binom{j}{p-\alpha}\binom{2}{\alpha}\binom{n-k-1}{p-1}\binom{c-1}{j}\frac{c!}{(c-j)!}\Psi(k-(j+1)s,c-1-j,s-1),$$

where again the Ψ-function can be explicitly obtained from (2.2) and (2.3) and where we define $\binom{n}{-1}$ as zero for any n and sums from $a > b$ to b are taken to be zero.

We now present an alternative method for solving the same problem, the resulting expression serving as the basis for all of Section 3. The assumption $c \ge 2$ will continue to be used. We work with the k covered points rather than the $n - k$ uncovered points, and we first consider the possible (unordered) sizes of groups of consecutive points which can comprise the k covered points; these groups are unconnected in the sense that there is at least one uncovered point between any two of them. For example, if $n = 10$, $k = 9$, and $s = 3$, then either the nine covered points are consecutive, or they consist of groups of six and three points separated by a point, or they consist of groups of five and four points separated by a point. These configurations are represented by (i) 9, (ii) 3, 6, (iii) 4, 5; configurations (ii) and (iii) include both the case where the smaller cluster is to the left of the larger one and the case where the opposite occurs. Note also that if $n = 11$ and k and s remain the same, then a fourth configuration (i.e., 3-3-3) is also possible, and configurations (ii) and (iii) now include cases where more than one point

separates the included groups. The number r of groups in a configuration may always be as small as one (i.e., the points are consecutive), but there are three constraints which may impose an upper limit on r. Obviously, we must have $r \le c$ because there cannot be more separate groups of included points than total clusters chosen. Also, $r \le [\frac{k}{s}]$ since the total number of covered points is k and no group of consecutive covered points may be smaller than s. Finally, there must be $r - 1$ uncovered points to separate the groups, so that $k + (r - 1) \le n$; hence $r \le n - k + 1$. It is easily verified that all three of these bounds are necessary, each one being smallest for certain cases.

Having established r, we now focus on the sizes $k_1, \ldots, k_r$ of the groups of covered points. In the above example (with $n = 11$) if $r = 1$ then $k_1 = 9$ and if $r = 3$ then $k_1 = k_2 = k_3 = 3$, but if $r = 2$ then we may have either $(k_1 = 4,\ k_2 = 5)$ or $(k_1 = 3,\ k_2 = 6)$. Note that we do require $s \le k_i \le k_{i+1}$ since we are unconcerned about order and because no group of covered points may be smaller than s.

For fixed r and $k_1, \ldots, k_r$, we wish to find the number $W_{k_1, \ldots, k_r}$ of ways that we can find unconnected groups of consecutive points of these sizes in the original chain. This depends on the multiplicities of the k_i's, which we call $m_1, m_2, \ldots, m_t$ so that $m_1 + m_2 + \ldots + m_t = r$. If all the k_i's are different, then $t = r$ and $m_1 = m_2 = \ldots = m_r = 1$. With this notation, there are $r! / \prod_{i=1}^{t} (m_i!)$ ways of ordering the r groups. Once the order is established, we first add one point between each of the groups to separate them; this accounts for $k + r - 1$ points. The remaining $S = n - k - r + 1$ points may be distributed in any manner into the $r + 1$ "slots" consisting of the two ends of the chain and the $r - 1$ gaps between the covered groups.

We are concerned only with the number of points put in each "slot", so the points may be considered indistinguishable. Since the number of ways to distribute M indistinguishable objects into N distinguishable slots is $\binom{M+N-1}{M}$, we write

$$(2.8) \qquad W_{k_1,\ldots,k_r} = \frac{r!}{\prod_{i=1}^{t} (m_i!)} \binom{S+r}{S} = \frac{r!}{\prod_{i=1}^{t} (m_i!)} \binom{S+r}{r} .$$

Since $\sum_{i=1}^{t} m_i = r$, we can rewrite the right hand side of (1.8) as a single multinomial coefficient:

$$(2.9) \qquad W_{k_1,\ldots,k_r} = \left[\begin{matrix} S+r \\ S, m_1, m_2, \ldots m_t \end{matrix} \right] .$$

We introduce the following notation: let $[[x]]$ be the smallest integer not less than x; hence $[[x]] = x$ if x is an integer and $[[x]] = [x] + 1$ otherwise. For each group k_i of consecutive covered points $(i = 1, 2, \ldots, r)$, we would like upper and lower limits for the number ℓ_i of <u>distinct</u> clusters which can cover this group. The upper limit is $k_i - s + 1$, the total number of clusters which exist if we consider the k_i points to comprise their own chain. The lower limit is $[[\frac{k_i}{s}]]$ since each cluster covers at most s new points. Each ℓ_i can vary within these limits independently of the others as long as $j = \sum_{i=1}^{r} \ell_i \leq c$, since the number of distinct clusters can never be greater than the total number of clusters. For each fixed pair (k_i, ℓ_i), we wish to find the number of sets of ℓ_i distinct clusters which cover the k_i consecutive points. As before, we can associate each cluster with its right endpoint and label all such possible right endpoints with the integers

$1, 2, \ldots, k_i - s + 1$. Again, clusters 1 and $k_i - s + 1$ must be included. Assuming $k_i > s$, these two clusters are distinct and $\ell_i \geq 2$, and we must choose $\ell_i - 2$ additional clusters without duplication such that no s consecutive points are unselected. But this means that we must divide the $k_i - s$ spaces between the cluster endpoints into $\ell_i - 1$ positive parts with no part greater than s. By definition, the number of ways to do this is $A_0(k_i - s, \ell_i - 1, s)$, where the A_0-function is given explicitly by (2.3). We have not covered the case $k_i = s$ and $\ell_i = 1$ (note that these two cases are equivalent); and since the answer for this case is one, we define $A_0(0, 0, s) = 1$ for any $s > 0$ to make our previous answer apply to this case as well. Furthermore, if all pairs (k_i, ℓ_i) are fixed $(i = 1, 2, \ldots, r)$, each group of k_i consecutive points can be covered by ℓ_i distinct clusters in the number of ways just obtained independently of which way is selected in any of the other groups. Therefore, the number of ways we can cover all the groups together by the required number of distinct clusters is the product $\prod_{i=1}^{r} A_0(k_i - s, \ell_i - 1, s)$.

Once we have obtained the specific $j = \sum_{i=1}^{r} \ell_i$ distinct clusters to be selected, we need the probability that all of these (and no others) are the ones that are chosen. It is clear that the probability of all selected clusters being among these j is $(\frac{j}{n-s+1})^c$. Given that all clusters are in this subset, the (conditional) distribution is multinomial with c trials and with each of the j clusters in question having probability $\frac{1}{j}$ at each trial. We want each of these clusters to be chosen at least once, meaning that the minimum frequency in this multinomial is at least one. This is given by

$$(2.10)\qquad I^{(j)}_{(1/j)}(1,\ c) = \frac{c!}{(c-j)!}\int_0^{1/j}\cdots\int_0^{1/j}\left(1-\sum_{i=1}^{j}x_i\right)\prod_{i=1}^{j}dx_i$$

where the I-function defines the desired condition on the minimum frequency and the notation for the I-function is consistent with [3], where the equality in (2.10) is derived.

Putting together the above entire discussion, we obtain

$$(2.11)\qquad P(k|n,c,s) = \sum_{r=1}^{\min(c,[\frac{k}{s}],n-k+1)}\ \sum_{\substack{k_1+\ldots+k_r=k\\ s\le k_i\le k_{i+1}\ (\forall i)}} W_{k_1,\ldots,k_r}\ \sum_{\substack{j=\ell_1+\ldots+\ell_r\le c\\ [[\frac{k_i}{s}]]\le \ell_i\le k_i-s+1\ (\forall i)}}\left[\prod_{i=1}^{r}A_0(k_i-s,\ell_i-1,s)\right]$$

$$\cdot\left(\frac{j}{n-s+1}\right)^c I^{(j)}_{(1/j)}(1,\ c)$$

where $W_{k_1,\ldots k_r}$ is given explicitly by (2.8) or (2.9) and the A_0-function by (2.3). It is easy to show using finite difference notation that the above I-function is related to a Stirling number of the second kind S^j_c since

$$(2.12)\qquad I^{(j)}_{(1/j)}(1,\ c) = \sum_{\alpha=0}^{j}(-1)^{\alpha}\binom{j}{j-\alpha}\left(\frac{j-\alpha}{j}\right)^c = (E-1)^j\left(\frac{x}{j}\right)^c\Big]_{x=0} = \Delta^j\left(\frac{x}{j}\right)^c\Big]_{x=0} = S^j_c\,\frac{j!}{j^c}\ .$$

Using (2.12), we rewrite (2.11) by

(2.13)

$$P(k|n,c,s) = \frac{1}{(n-s+1)^c} \sum_{r=1}^{\min(c,[\frac{k}{s}],n-k+1)} \sum_{\substack{k_1+\ldots+k_r=k \\ s\le k_i\le k_{i+1}\ (\forall i)}} W_{k_1,\ldots,k_r} \sum_{\substack{j=\ell_1+\ldots+\ell_r\le c \\ [[\frac{k_i}{s}]]\le \ell_i\le k_i-s+1\ (\forall i)}}$$

$$\cdot \Big[\prod_{i=1}^{r} A_0(k_i-s,\ \ell_i-1,s)\Big]\ j!\ S_c^j\ .$$

This method has the advantage that we can get the results for $k = n$ and $k = s$ without treating these as special cases. For complete coverage (i.e., $k = n$), we get $r = 1$, $k_1 = k = n$, $W_{k_1,\ldots,k_r} = W_n = 1$, and $\ell_1+\ldots+\ell_r = \ell_1 = j$. Hence, (2.13) reduces to

(2.14)
$$P(n|n,c,s) = \frac{1}{(n-s+1)^c} \sum_{j=[[\frac{n}{s}]]}^{\min(c,n-s+1)} A_0(n-s,\ j-1,\ s)\ j!\ S_c^j\ .$$

As an illustration we consider the calculations for $P(k|6,4,3)$ for $k = 3,4,5$, and 6. Using either (2.13) or a combination of (2.7), (2.1), and the special result for $k = s$ we obtain

(2.15)
$$P(3|6,4,3) = \frac{4}{256},\ P(4|6,4,3) = \frac{42}{256},\ P(5|6,4,3) = \frac{100}{256},\ P(6|6,4,3) = \frac{110}{256}\ .$$

These results yield an expectation $E(K) = 335/64 = 5\frac{15}{64}$, which agrees with a result obtained in [2]. The variance is easily seen to be 2,463/4,096.

B. Circular Chain, With Replacement

The basic technique for our alternative method from Case A works here as well for $k < n$ but not for $k = n$, i.e., the case of complete coverage. Assuming $k < n$, we make the following changes in (2.13):

(i) The factor $(\frac{j}{n-s+1})^c$ in the discussion leading to (2.10) becomes $(\frac{j}{n})^c$ since there are now n clusters to choose from. This carries through to (2.11), so that the outside factor in (2.13) and (2.14) becomes $\frac{1}{n^c}$.

(ii) After fixing r, we now need r uncovered points to separate the groups rather than just $r - 1$. Hence $k + r \leq n$ and $r \leq n - k$; thus one of the upper limits for r in (2.13) must be decreased by one. This is one minor place where the case $n = k$ fails since the new upper limit on r would then give an empty sum (and a zero probability), whereas we clearly need to use $r = 1$ in this case.

(iii) The entire computation for $W_{k_1,\ldots,k_r}$ must be altered, and we do this below.

Before doing this, we remark that all terms and factors to the right of $W_{k_1,\ldots k_r}$ in (2.13) were validly computed independently of linearity of the chain. Hence (for $k < n$) (i), (ii), and (iii) are the only required modifications of (2.13). This is the major failure of our method for $k = n$, namely that the A_0-function in (2.13) represents the number of ways to cover a linear group of consecutive points by a specified number of distinct clusters, and for $k = n$ our group of consecutive points is the entire (circular) chain.

Let $M_{k_1,\ldots,k_r}$ denote the new $W_{k_1,\ldots,k_r}$, i.e., the number of ways we can find r separated groups of consecutive points with sizes $k_1,\ldots,k_r$. Unlike Case A, we take $k_1,\ldots,k_r$ to be an ordered partition of k since we

may not get the same answer for each of the $r!/\prod_{i=1}^{t}(m_i!)$ orderings of the k_i's as we did before. Note that a circular re-ordering of the numbers in the partition (e.g., placing k_1 at the end) may be considered different even though the groups of points will be placed on a circle because, for example, different numbers of uncovered points will be inserted after the first group and after the last group. For any particular (ordered) partition $k_1,\ldots,k_r$, we first add r uncovered points to separate the groups, one after each group. The remaining $T = n - k - r$ indistinguishable points are then inserted into the r distinguishable gaps which already contain one point each, and this can be done in $\binom{T+r-1}{T}$ ways. For a fixed way of inserting these points, the configuration can be rotated around the points of the circle. That is, each of the n points can be used as the first point in the first of our r groups of consecutive covered points, which seemingly results in an additional factor of n. It would then appear from the above that

$$M_{k_1,\ldots,k_r} = n\binom{T+r-1}{T}, \tag{2.16}$$

but there may be some duplications to eliminate.

We define an ordered sequence of numbers of finite length r to be periodic with period of size σ if it consists only of <u>full</u> repetitions of an ordered sequence of numbers of length σ. We can always take $\sigma = r$ since the whole sequence is also regarded as a repetition. Hence every finite sequence of numbers is periodic for some σ. Suppose our partition $k_1,\ldots,k_r$ is periodic with smallest period σ. Then $\rho = \frac{r}{\sigma}$ is the number of repetitions of the period. Suppose further that $J = \frac{T}{\rho}$ is an integer. Then the T indistinguishable points above may be inserted identically relative to the cor-

responding elements of the different periods, with J points within (or following the last element of) each period. The number of ways that this can happen is the number of ways that J of the indistinguishable points can be distributed among the σ gaps following the elements of any period, which is $\binom{J+\sigma-1}{J}$. When this happens, we cannot rotate our configuration to begin at all n points but only at $\frac{n}{\rho}$ of them, the configuration being the same for two beginning points which are equal (mod ρ). Note that $\frac{n}{\rho} = \frac{(n-k-r)+k+r}{\rho} = \frac{T}{\rho} + \frac{k}{\rho} + \frac{r}{\rho} = J + (k_1+k_2+\ldots+k_\sigma) + \sigma$ is an integer, providing a partial check on the above argument. We conclude that if $T \equiv 0 \pmod{\rho}$,

$$(2.17) \qquad M_{k_1,\ldots,k_r} = n\left\{\binom{T+r-1}{T} - \binom{J+\sigma-1}{J}\right\} + \frac{n}{\rho}\binom{J+\sigma-1}{J}$$

$$= n\left\{\binom{T+r-1}{T} - \frac{(\rho-1)}{\rho}\binom{J+\sigma-1}{J}\right\}$$

For $T \not\equiv 0 \pmod{\rho}$, there are no duplications, and (2.16) does hold. Note that for $\rho = 1$ (i.e., no non-trivial period exists), $\frac{T}{\rho}$ is an integer and (2.17) applies. But in that case there are no duplications, and (2.17) reduces to (2.16) as we would expect.

We now let $N_{k_1,\ldots,k_r} = \frac{1}{n} M_{k_1,\ldots,k_r}$. This enables us to remove the factor n in (2.16) and (2.17) and cancel it against the n^c in the denominator. Thus, for $k < n$,

$$(2.18) \qquad P(k|n,c,s) = \frac{1}{n^{c-1}} \sum_{r=1}^{\min(c,[\frac{k}{s}],n-k)} \sum_{\substack{k_1+\ldots+k_r=k \\ s\le k_i\ (\forall i)}} N_{k_1,\ldots,k_r} \cdot \Big\{\prod_{i=1}^{r} A_0(k_i-s,\ell_i-1,s)\Big\} j!\, S_c^j$$

$$\sum_{\substack{j=\ell_1+\ldots+\ell_r \le c \\ [[\frac{k_i}{s}]] \le \ell_i \le k_i-s+1\ (\forall i)}}$$

where $N_{k_1,\ldots,k_r} = \frac{1}{n} M_{k_1,\ldots,k_r}$ and the M-value is given by either (2.16) or (2.17). The decision as to which of (2.16) or (2.17) to use depends on ρ, which in turn depends on the order of the k_i's, which justifies our need for $k_1,\ldots,k_r$ to be ordered. Formula (2.18) indicates this by eliminating the condition $k_i \leq k_{i+1}$ found in (2.13), and this eliminates from (2.16) and (2.17) the factor for ordering in (2.8) and (2.9). Note that if $m_1,\ldots,m_t$ are relatively prime, and in particular if any $m_i = 1$ for $1 \leq i \leq t$, we will always have $\rho = 1$ so that both (2.16) and (2.17) are correct independently of the order of the k_i's, and (2.18) can be adjusted by the ordering factor $r!/\prod_{i=1}^{t}(m_i!)$ to be a sum on unordered partitions of k so that it parallels (2.13) more closely.

To solve the complete coverage case $(k = n)$, we first point out that the probability of any event describing the number of points in the union or intersection of the clusters is independent of the choice of the first cluster by symmetry of the circle. Thus, we can use the first cluster to break the circle into a linear chain (causing the positions of the remaining clusters relative to the first cluster to be relevant), and thereafter ignore the first cluster completely. We now consider ourselves to be sampling only $c - 1$ clusters, and our new sample space has only n^{c-1} elements. Incidentally, this helps justify the fact that the power of n in the denominator of (2.18) is $c - 1$ rather than c. If, as usual, we associate each cluster with its right endpoint, we arbitrarily break up the circle (including the spaces between the points) into a linear chain which ends with the right endpoint of our original cluster. Hence, our chain begins with a space. By choosing $c - 1$ more points we are partitioning the n spaces into c non-negative parts,

and it is not hard to see that complete coverage results when no part is greater than s. Furthermore, our partition may not begin with a zero since the chain begins with a space. In terms of our new sample space, a terminal zero in our partition has a different meaning from a zero elsewhere. For a zero in the middle of the partition means that some point (i.e., cluster location) was chosen twice among our $c - 1$ clusters, whereas a terminal zero means that the first point was chosen again later, but still only once in terms of our sample space defined by $c - 1$ clusters since the first cluster is not part of this space. In general, by the same reasoning, a set of z consecutive zeros in the middle of our partition indicates a point chosen with multiplicity z, but if the zeros are at the end of the partition the multiplicity is only $z - 1$ in terms of our sample space.

If we define $A(n,c,s)$ to be the number of ordered partitions of n into c non-negative parts with each part not greater than s (so that the A-function is the sum of all the subscripted A-functions with the same arguments, which were defined between (2.1) and (2.2)), then

$$B(n,c,s) = A(n,c,s) - A(n,c-1,s) \tag{2.19}$$

is the number of partitions included in $A(n,c,s)$ with the first part positive. If $B_0(n,c,s)$ is the number of such partitions with no zero parts, then by definition

$$B_0(n,c,s) = A_0(n,c,s) \tag{2.20}$$

where the A_0-function is given by (2.3) and defined prior to (2.2). Since in terms of multiplicities we wish to group the cases where the only zero is terminal together with those with no zeros at all, we denote the total number of cases in these two categories by $B_0^*(n,c,s)$. If the only zero is terminal then n is really being partitioned into $c - 1$ positive parts, so that

$$(2.21) \qquad B_0^*(n,c,s) = B_0(n,c,s) + B_0(n,c-1,s) = A_0(n,c,s) + A_0(n,c-1,s).$$

We now designate by $B_1^*(n,c,s)$ the number of allowable partitions with exactly one non-terminal zero (and clarify that only one of a string of successive zeros at the end of our partition is to be regarded as terminal); $B_2^*(n,c,s)$ denotes the number of partitions with two successive non-terminal zeros and no others; and other subscripted B^*-functions are defined according to the structure of the non-terminal zeros in the manner of the subscripted A-functions defined prior to (2.2). The subscripted B^*-functions divide our set of partitions into subsets where the multiplicities which determine in how many orders each partition can be obtained remain constant within each subset. Hence, for complete coverage we obtain

(2.22)

$$P(n|n,c,s) = \frac{(c-1)!}{n^{c-1}} \left(B_0^*(n,c,s) + \frac{B_1^*(n,c,s)}{2!} + \left\{\frac{B_2^*(n,c,s)}{3!} + \frac{B_{1,1}^*(n,c,s)}{(2!)^2}\right\}\right.$$
$$\left. + \left\{\frac{B_3^*(n,c,s)}{4!} + \frac{B_{2,1}^*(n,c,s)}{3!\,2!} + \frac{B_{1,1,1}^*(n,c,s)}{(2!)^3}\right\} + \ldots\right).$$

We note that a partition included in $B_1^*(n,c,s)$ can be obtained by adding a zero to a partition included in $B_0^*(n,c-1,s)$. However, this zero

may not be placed at the beginning of the partition, and placing it at the end of the partition creates a non-terminal zero only if a terminal zero already exists, in which case we can equivalently place it just prior to that zero (and rule out the case where the zero is placed at the end). Hence, the zero can be placed only in the $c - 2$ gaps located between two of the $c - 1$ elements of the partition. By counting the number of arrangements for the zeros, we can similarly express all the subscripted B^*-functions in terms of the B_0^*-functions, e.g.:

$$B_1^*(n,c,s) = \binom{c-2}{1} B_0^*(n,c-1,s)$$

$$B_2^*(n,c,s) = \binom{c-3}{1} B_0^*(n,c-2,s)$$

$$B_{1,1}^*(n,c,s) = \binom{c-3}{2} B_0^*(n,c-2,s) \tag{2.23}$$

$$B_3^*(n,c,s) = \binom{c-4}{1} B_0^*(n,c-3,s)$$

$$B_{2,1}^*(n,c,s) = 2\binom{c-4}{2} B_0^*(n,c-3,s)$$

$$B_{1,1,1}^*(n,c,s) = \binom{c-4}{3} B_0^*(n,c-3,s), \text{ etc.}$$

In general, if $a_1, a_2, \ldots, a_r$ are the distinct integers in the subscripts of B^* and if each a_i occurs m_i times $(i = 1, 2, \ldots, r)$, then we set $z = \sum_{i=1}^{r} a_i m_i$ and $y = \sum_{i=1}^{r} m_i$ and write

$$B^*_{a_1^{m_1}, a_2^{m_2}, \ldots, a_r^{m_r}}(n,c,s) = \left[{}^{y}_{m_1,m_2,\ldots,m_r}\right]\binom{c-z-1}{y} B_0^*(n,c-z,s) \tag{2.24}$$

$$= \left[{}^{c-z-1}_{m_1,m_2,\ldots,m_r,c-z-1-y}\right] B_0^*(n,c-z,s).$$

We now write (2.22) in terms of B_0^*-functions only:

(2.25)

$$P(n|n,c,s) = \frac{(c-1)!}{n^{c-1}} \left(B_0^*(n,c,s)\{1\} + B_0^*(n,c-1,s)\{\frac{\binom{c-2}{1}}{2!}\} + B_0^*(n,c-2,s)\{\frac{\binom{c-3}{1}}{3!} + \frac{\binom{c-3}{2}}{(2!)^2}\}\right.$$

$$\left. + B_0^*(n,c-3,s)\{\frac{\binom{c-4}{1}}{4!} + \frac{2\binom{c-4}{2}}{2!\,3!} + \frac{\binom{c-4}{3}}{(2!)^3}\} + \ldots\right)$$

where the expressions in braces are a sequence of polynomials in c of increasing degree, and we denote by $Q_i(c)$ the polynomial of degree i so obtained ($i = 0, 1, 2\ldots$). Analogous to (2.2), the expressions $(c-1)!\,Q_i(c)$ are integers which depend only on c and have some interest per se. Using (2.21), we write (2.25) in terms of A_0-functions only:

$$(2.26) \qquad P(n|n,c,s) = \frac{(c-1)!}{n^{c-1}}\left(A_0(n,c,s)\{Q_0(c)\} + A_0(n,c-1,s)\{Q_0(c) + Q_1(c)\}\right.$$

$$\left. + A_0(n,c-2,s)\{Q_1(c) + Q_2(c)\} + A_0(n,c-3,s)\{Q_2(c) + Q_3(c)\} + \ldots\right)$$

where (2.26) includes only part of the last term we wrote out in (2.25). Again, the A_0-functions are given explicitly by (2.3). (See insert following page.)

C. Linear Chain, Without Replacement

This is very similar to Case A and we use the symbols in (2.13). Here every cluster must be distinct but of course for $s > 1$ any pair of clusters can have points in common. Hence $j = \sum_{i=1}^{r} \ell_i = c$ because j is the number of distinct clusters. Since $[[\frac{k_i}{s}]] \le \ell_i \le k_i - s + 1$, we sum these inequalities

Insert at end of 2.B.

We illustrate the results of this section by computing $P(k|6,4,3)$ for $k = 3,4,5,$ and 6. Using (2.18) we obtain $P(3|6,4,3) = \frac{1}{216}$, $P(4|6,4,3) = \frac{14}{216}$, and $P(5|6,4,3) = \frac{50}{216}$. We then use (2.25) to obtain $P(6|6,4,3) = \frac{151}{216}$. Note that we use only three terms of the "infinite" sum in (2.25) and that this sum is finite in general since $B_0^*(n,c-j,s) = 0$ when j is large enough that $s(c-j) < n$. The above probabilities yield $E(K) = \frac{45}{8}$ (which checks a result from [2]) and $Var(K) = \frac{677}{1,728}$. (Compare all of these numerical results with (2.15) and the two lines after it.)

over $i = 1, 2, \ldots, r$ to obtain $\sum_{i=1}^{r} [[\frac{k_i}{s}]] \le c \le \sum_{i=1}^{r} (k_i - s + 1) = k - r(s-1)$, the equality being due to the fact that $\sum_{i=1}^{r} k_i = k$. Then $r \le [\frac{k-c}{s-1}]$ is a new upper bound on r, and this renders unnecessary the previous bound $r \le [\frac{k}{s}]$ since $\frac{k-c}{s-1} \le \frac{k}{s} \Leftrightarrow s(k-c) \le k(s-1) \Leftrightarrow k \le cs$, which is obviously always true. Also, the smallest number of points that can be covered by c clusters is $s + c - 1$, the size of the smallest linear chain which contains c possible clusters. Hence $k \ge s + c - 1$, and we should get $P(k|n,c,s) = 0$ for $k < s + c - 1$. But in this case $[\frac{k-c}{s-1}] = 0$ so that due to our new upper bound on r, we are summing on r from one to zero; hence $P(k|n,c,s) = 0$ as desired. To illustrate the restriction on the k_i's imposed by the inequality $\sum_{i=1}^{r} [[\frac{k_i}{s}]] \le c$ that was obtained above, we take $n = 11$, $c = s = 3$, and $k = 8$. Then $r = 2$ is permissible, a fact which can be verified either by checking the bounds on r or by noting that if $k_1 = 3$ and $k_2 = 5$, we can cover this configuration of points with the three allotted clusters. But $k_1 = k_2 = 4$ would require four clusters to cover the points and cannot be an allowable partition of k, a fact obtainable by our formulas only through the new restriction on the k_i's since $[[\frac{k_1}{s}]] + [[\frac{k_2}{s}]] = 2 + 2 = 4 > 3 = c$. We note in passing that without the double brackets we would not have any restriction at all since $\sum_{i=1}^{r} \frac{k_i}{s} = \frac{k}{s}$, which is clearly never greater than c. Finally, due to the fact that no cluster may be repeated, we can disregard the order in which the clusters are chosen and define a sample space of size $\binom{n-s+1}{c}$, each element corresponding to a different possible set of the c clusters which are chosen. Hence, each particular set of $j = c$ clusters has probability $1/\binom{n-s+1}{c}$ of being chosen, and this quantity replaces the $(\frac{j}{n-s+1})^c$ together with the I-function in (2.11) or, analogously, the $\frac{j!}{(n-s+1)^c}$ together with the Stirling number in (2.13). Putting together all of the above facts,

we get for the linear chain without replacement

(2.27)

$$P(k|n,c,s) = \frac{1}{\binom{n-s+1}{c}} \sum_{r=1}^{\min(c,[\frac{k-c}{s-1}],n-k+1)} \sum_{k_1+\ldots+k_r=k} W_{k_1,\ldots,k_r} \sum_{\substack{s\le k_i\le k_{i+1}\ (\forall i) \\ \sum_{i=1}^{r}[[\frac{k_i}{s}]]\le c}} \sum_{\substack{\ell_1+\ldots+\ell_r=c \\ [[\frac{k_i}{s}]]\le \ell_i \le k_i-s+1\ (\forall i)}} \cdot \prod_{i+1}^{r} A_0(k_i-s,\ell_i-1,s),$$

where $W_{k_1,\ldots,k_r}$ is given by (2.9), which remains valid. Note from (2.27) that the k_i's again represent unordered partitions of k. (See insert following page.)

D. Circular Chain, Without Replacement

For the complete coverage case $(k = n)$, we break the circle after the first cluster is chosen, as in Case B, so that we have a linear collection of spaces and points which begins with a space and ends with a point. We define a new sample space of size $\binom{n-1}{c-1}$ rather than the original $\binom{n}{c}$ by ignoring both the first point (as in Case B) and the order in which the clusters were chosen (as in Case C). Since repetition is not allowed, the number of ways we can choose a set of clusters which will completely cover the points is the same as the number of ways we can choose $c - 1$ points which break up the n spaces into c positive parts, each at most s. But this is just $A_0(n,c,s)$, which retains its old definition and is given by (2.3). Hence for $k = n$,

(2.28) $$P(n|n,c,s) = \frac{1}{\binom{n-1}{c-1}} A_0(n,c,s).$$

Insert at end of 2.C.

Using (2.27), we obtain $P(6|6,4,3) = 1$; hence the example used in A and B is not very illuminating here, and we instead use (2.27) to write the results for $n = 10$, $c = 4$, $s = 3$. They are $P(6|10,4,3) = \frac{5}{70}$, $P(7|10,4,3) = \frac{12}{70}$, $P(8|10,4,3) = \frac{27}{70}$, $P(9|10,4,3) = \frac{20}{70}$, and $P(10|10,4,3) = \frac{6}{70}$. Hence $E(K) = \frac{57}{7}$ and $Var(K) = \frac{261}{245}$.

For $k < n$, we use the entire argument from Case C to make the same changes in (2.18) that were made in (2.13) to arrive at (2.27) except that the stipulation $k \geq s+c-1$, which had no bearing on the written form of (2.27) anyway, becomes $k \geq \min(n, s+c-1)$ since we now allow $c > n-s+1$. We do not repeat this argument, but we make one additional change. The factor $\frac{1}{n^{c-1}} N_{k_1,\ldots,k_r}$ in (2.18) originally came from $\frac{1}{n^c} M_{k_1,\ldots,k_r} = \frac{n}{n^c} N_{k_1,\ldots,k_r}$. Here the numerator remains the same, but the n^c in the denominator becomes $\binom{n}{c}$ just as the $(n-s+1)^c$ in (2.13) became $\binom{n-s+1}{c}$ in (2.27) since we are now choosing c distinct clusters from the total set of clusters. Thus, the appropriate version of (2.18) is

(2.29)

$$P(k|n,c,s) = \frac{n}{\binom{n}{c}} \sum_{r=1}^{\min(c,[\frac{k-c}{s-1}],n-k)} \sum_{\substack{k_1+\ldots+k_r=k \\ s\leq k_i \ (\forall i)}} N_{k_1,\ldots,k_r} \sum_{\substack{\ell_1+\ldots+\ell_r=c \\ [[\frac{k_i}{s}]] \leq \ell_i \leq k_i-s+1 \ (\forall i)}} \prod_{i=1}^{r} A_o(k_i-s, \ell_i-1, s)$$

$$\sum_{i=1}^{r} [[\frac{k_i}{s}]] \leq c$$

where $N_{k_1,\ldots,k_r} = \frac{1}{n} M_{k_1,\ldots,k_r}$ and the M-function is given by either (2.16) or (2.17), both of which remain valid. Note that the k_i's are again ordered.

We might ask why the outside denominator of (2.29) is not $\binom{n-1}{c-1}$, the number of points in the applicable smaller sample space used to obtain (2.28), which would create a good parallel to the fact that (2.18) and (2.25) have the same outside denominator. The answer is that this parallel really does exist since $\frac{n}{\binom{n}{c}} = \frac{c}{\binom{n-1}{c-1}}$.

Using (2.29), we obtain $P(6|10,4,3) = \frac{10}{210}$, $P(7|10,4,3) = \frac{30}{210}$, $P(8|10,4,3) = \frac{75}{210}$, $P(9|10,4,3) = \frac{70}{210}$; (2.28) yields $P(10|10,4,3) = \frac{25}{210}$. Thus $E(K) = \frac{25}{3}$ and $Var(K) = \frac{65}{63}$. Thus, at least for the particular illustrations we have used, the union has a larger expected value and a smaller variance in the circular case than in the linear case whether we work with or without replacement. We conjecture from intuition that this is true in general, but that the means and variances for the circular and linear cases are asymptotically equal as $n \to \infty$; we have not found a proof for our conjectures.

E. General Comments for A, B, C, and D.

We emphasize the obvious fact that though we have not found a method to sum the probabilities to obtain $E(K)$ or higher order moments of K, all of these moments can be found for any specific values of n, c, and s by numerical computation after all the individual probabilities are calculated from the appropriate formulas given above. Thus, we have generalized the work done in [2].

3. PROBABILITY THAT THE INTERSECTION K HAS EXACTLY k POINTS.

Before considering each case separately, we note that the result is zero in all cases if $k > s$. Hence, we consider only $0 \le k \le s$. The notation $P(k|n,c,s)$ continues to represent $P\{K = k|n,c,s\}$ with the order of n, c, and s being relevant as before.

A. Linear Chain, With Replacement.

If $0 < k \le s$, then the maximum distance between right cluster endpoints is exactly $s - k$ (i.e., the right endpoints of the c clusters fall within a succession of $s - k + 1$ possible endpoints in the chain and not in any proper contiguous subset of these endpoints). The probability that the above property will be satisfied and that the intersection will be some particular set of k successive points is therefore

$$(3.1) \qquad Q(k|n,c,s) = \frac{(s-k+1)^c - 2(s-k)^c + (s-k-1)^c \delta'_{sk}}{(n-s+1)^c}$$

where $\delta'_{sk} = 0$ if $s = k$ and $\delta'_{sk} = 1$ if $s \neq k$. To have an intersection of any size k $(0 < k \le s)$, we see from the above argument that $n \ge 2s - k$. Furthermore, there are $n + 1 - 2s + k$ possible positions for the k successive points in the intersection (assuming this quantity is positive). Hence for $n \ge 2s - k$,

$$(3.2) \qquad P(k|n,c,s) = (n+1-2s+k)\, Q(k|n,c,s)$$

$$= \frac{n+1-2s+k}{(n-s+1)^c} \{(s-k+1)^c - 2(s-k)^c + (s-k-1)^c \delta'_{sk}\} ;$$

the result is zero for $n \le 2s - k - 1$. Note that (3.2) corresponds to this result (and is valid) for $n = 2s - k - 1$, except that this is meaningless for $k = s$ since we require $n \ge s$. Thus (3.2) is valid for every $1 \le k \le s$ if $n \ge 2s - 2$.

The following identity, which holds for any positive integer c and non-negative integers i and S is used below:

$$(3.3) \qquad \sum_{\alpha=0}^{S} \alpha^i\{(\alpha+1)^c - 2\alpha^c + (\alpha-1)^c\delta'_{\alpha o}\} = S^i(S+1)^c - (S+1)^i S^c + 2\sum_{j=1}^{S}\sum_{\beta=1}^{[\frac{i}{2}]}\binom{i}{2\beta} j^{c+i-2\beta}.$$

The proof of (3.3) uses the fact that the coefficient of α^c for each α is $(\alpha-1)^i + (\alpha+1)^i - 2\alpha^i$; we omit further details.

For the special case $k = 0$ we set $\alpha = s - k$ and use (3.3) with $i = 0,1$ and $S = s - 1$, and obtain for $n \ge 2s - 2$

$$(3.4) \qquad P(0|n,c,s) = 1 - \sum_{k=1}^{s} P(k|n,c,s)$$

$$= 1 - \frac{1}{(n-s+1)^c}\sum_{\alpha=0}^{s-1}(n+1-s-\alpha)\{(\alpha+1)^c - 2\alpha^c + (\alpha-1)^c\delta'_{\alpha o}\}$$

$$= 1 - \{\frac{(n-2s+2)s^c - (n-2s+1)\ (s-1)^c}{(n-s+1)^c}\}\ .$$

When $n < 2s$, $P(0|n,c,s)$ must be zero for any c since the s^{th} point from each end (and all points between these two) is then contained in every possible cluster. As a partial check on (3.4) we note that it does yield zero for $n = 2s - 1$ and $n = 2s - 2$.

By definition, $E(K|n,c,s) = \sum_{k=1}^{s} kP(k|n,c,s) = \sum_{\alpha=0}^{s-1}(s-\alpha)P(s-\alpha|n,c,s)$. But by the remark after (3.2), all terms in this sum for which $k \le 2s - n - 1$ (i.e., for which $\alpha \ge n - s + 1$) vanish. Hence the lower limit on k is taken to be $\max(1,2s-n)$ and the upper limit on α is then $S = \min(s-1,n-s)$. However, the remark after (3.2) also shows that we may choose not to omit the term $\alpha = n-s+1$ when we sum the probabilities given by (3.2). Thus the computations below are also valid if we use $S = \min(s-1, n-s+1)$ and the final formulas are identical. In either case, we obtain using (3.3) with $i = 0$, $i = 1$, and $i = 2$,

$$(3.5) \qquad E(K|n,c,s) = \frac{1}{(n-s+1)^c} \sum_{\alpha=0}^{S}(s-\alpha)(n+1-s-\alpha)\{(\alpha+1)^c - 2\alpha^c + (\alpha-1)^c\delta'_{\alpha o}\}$$

$$= \frac{1}{(n-s+1)^c}\{(S^2 - s^2 - (n+1)(S-s))(S+1)^c$$

$$- (S^2 - s^2 - (n+1)(S-s) + 2S - n)S^c + 2\sum_{j=1}^{S} j^c\}$$

$$= \begin{cases} \dfrac{1}{(n-s+1)^c}\{2\sum_{j=1}^{s-1} j^c + (n-2s+2)s^c\} & \text{for } n \ge 2s-2 \\ 2s - 2 - n + 2\sum_{j=1}^{n-s+1}\left(\dfrac{j}{n-s+1}\right)^c & \text{for } s \le n \le 2s-1 . \end{cases}$$

Note that the two final versions of (3.5) agree when $n = 2s - 1$ and when $n = 2s - 2$, a fact which also follows by noting that these are the values of n that are obtained by equating the components within each of the min functions which define S. By arbitrarily using the first version for

$n = 2s - 1$ and the second definition of S, we combine both versions into

$$(3.6)\quad E(K|n,c,s) = 2\sum_{j=1}^{S}\left(\frac{j}{n-s+1}\right)^c + |n-2s+2|\ \min\left\{1, \left(\frac{s}{n-s+1}\right)^c\right\}.$$

By using the same technique as above, retaining both definitions of S, and applying (3.3) with $i = 0, 1, 2, 3$, we obtain

$$(3.7)\quad E(K^2|n,c,s) = \frac{1}{(n-s+1)^c}\sum_{\alpha=0}^{S}(s-\alpha)^2(n-s+1-\alpha)\{(\alpha+1)^c - 2\alpha^c + (\alpha-1)^c\delta'_{\alpha o}\}$$

$$= \frac{1}{(n-s+1)^c}\{(s^2(n-s+1) - sS(2n-s+2) + S^2(n+s+1) - S^3)(S+1)^c$$

$$- (s^2(n-s+1) - s(S+1)(2n-s+2) + (S+1)^2(n+s+1) - (S+1)^3)S^c$$

$$+ 2(n+s+1)\sum_{j=1}^{S}j^c - 6\sum_{j=1}^{S}j^{c+1}\}$$

$$= \begin{cases} \dfrac{(n-2s+2)s^c + 2(n+s+1)\sum_{j=1}^{s-1}j^c - 6\sum_{j=1}^{s-1}j^{c+1}}{(n-s+1)^c} & \text{for } n \geq 2s-2 \\ (n-2s+2)^2 + \dfrac{2(n+1+s)\sum_{j=1}^{n-s+1}j^c - 6\sum_{j=1}^{n-s+1}j^{c+1}}{(n-s+1)^c} & \text{for } s \leq n \leq 2s-1. \end{cases}$$

These two final expressions are again equal for $n = 2s-1$ and $n = 2s-2$, and the reasons for this are the same as before. No natural way seems to exist to combine the above in the manner of (3.6). We now can, of course, write

(3.8) $\quad \text{Var}\,(K|n,c,s) = E(K^2|n,c,s) - E^2(K|n,c,s)$

where the right side of (3.8) can be evaluated using (3.6) and (3.7). We remark that the first of the two final versions of (3.5) checks a result obtained in [2], but the second version represents a new result, as do (3.6) and (3.7). Also, for $c = 1$ we have $K = s$ with probability one for all values of n and s, and in this case it can be verified that both versions of (3.5) (hence also (3.6)) yield s, both forms of (3.7) give s^2, and (3.8) therefore is zero.

We now derive an alternative expression for $P(k|n,c,s)$ in Case A. For $k = s$, the first cluster may be arbitrarily selected and all the others must coincide. Hence the result is $(n-s+1)^{-(c-1)}$; in particular, this yields $P(s|n,1,s) = 1$ for any values of n and s (and the cases $c = 1$ and $k = s$ are solved). For $0 < k < s$ and $c > 1$, we obtain as before the result zero for $n \le 2s - k - 1$ and the fact that there are $n+1-2s+k$ possible positions for the k successive points in the intersection if $n \ge 2s-k$. For any particular one of these sets of k points, we need the two extreme clusters which have these k points in common. The remaining clusters must all be from the set of $s-k+1$ clusters whose right endpoints coincide with or lie between those of the two clusters already chosen; hence the maximum number of distinct clusters different from the two extremes is $s-k-1$. Suppose we have i from this set. These can be chosen in $\binom{s-k-1}{i}$ ways, and for each such choice there corresponds a particular set of $i+2$ distinct clusters which must consist exactly of all those clusters

selected at least once and no others. The probability that this happens, which is obtained by the same reasoning that led to (2.11), (2.12), and (2.13), is $(\frac{i+2}{n-s+1})^c \, I^{(i+2)}_{(\frac{1}{i+2})}(1,c) = \frac{(i+2)!}{(n-s+1)^c} S_c^{i+2}$. The above I-function is given in terms of an (i+2)-fold integral by (2.10). Hence for $0 < k < s$, $c > 1$, and $n \geq 2s-k$, we obtain (defining $S_\alpha^\beta = 0$ if $\beta > \alpha$)

$$(3.9) \qquad P(k|n,c,s) = \frac{n+1-2s+k}{(n-s+1)^c} \sum_{i=0}^{s-k-1} \binom{s-k-1}{i} (i+2)^c \, I^{(i+2)}_{(\frac{1}{i+2})}(1,c)$$

$$= \frac{n+1-2s+k}{(n-s+1)^c} \sum_{i=0}^{s-k-1} \binom{s-k-1}{i} (i+2)! \, S_c^{i+2} .$$

For the special case $k = 0$ we obtain by subtraction for $c > 1$ and $n \geq 2s - 2$

$$(3.10) \quad P(0|n,c,s) = 1 - \frac{1}{(n-s+1)^{c-1}} - \sum_{k=1}^{s-1} \frac{n+1-2s+k}{(n-s+1)^c} \sum_{i=0}^{s-k-1} \binom{s-k-1}{i} (i+2)! S_c^{i+2} ;$$

the result is zero as before for $n < 2s$.

All the intersection formulas are related to union formulas, and this will be discussed later in Section 4. (See insert following page.)

B. Circular Chain, With Replacement.

For $n \geq 2s - 1$, the same method used in Case A also applies here. Replacing both $n + 1 - 2s + k$ and $n - s + 1$ in (3.2) by n to account for circularity, we obtain for $0 < k \leq s$ by the same reasoning

$$(3.11) \qquad P(k|n,c,s) = \frac{1}{n^{c-1}} \{(s-k+1)^c - 2(s-k)^c + (s-k-1)^c \, \delta'_{sk}\} .$$

Insert at end of 3.A.

In fact, a numerical example can be obtained for this section by merely applying the 1-1 correspondence between union and intersection for $n = 6$, $c = 4$, $s = 3$ that we will establish is Section 4 to the numerical probabilities obtained at the end of Section 2, Case A. Moreover, by using the remarks following (4.5) as well as (4.1) and (4.3), we can obtain the mean and the variance as well as the individual probabilities without doing any additional work. All of these quantities turn out to be consistent with (3.2), (3.4), (3.6), and (3.8).

Using (3.3) with $i = 0, 1, 2, 3$, we obtain for $n \geq 2s - 1$,

$$(3.12) \qquad P(0|n,c,s) = 1 - \{ \frac{s^c-(s-1)^c}{n^{c-1}} \} ,$$

$$(3.13) \qquad E(K|n,c,s) = \frac{1}{n^{c-1}} \{s[s^c-(s-1)^c] - [(s-1)s^c-s(s-1)^c]\} = \frac{s^c}{n^{c-1}} = n(\frac{s}{n})^c,$$

$$(3.14) \qquad E(K^2|n,c,s) = \frac{1}{n^{c-1}} (2 \sum_{j=1}^{s-1} j^c+s^c) = \frac{1}{n^{c-1}} (2 \sum_{j=1}^{s} j^c - s^c), \quad \text{and}$$

$$(3.15) \qquad Var(K|n,c,s) = EK^2 - (EK)^2 = \frac{n^{c-1}(2 \sum_{j=1}^{s-1} j^c + s^c) - s^{2c}}{n^{2c-2}} .$$

Note that (3.13) also checks a result from [2]. For $c = 1$ (when $K = s$ with probability one), we can again verify that (3.13) yields s, (3.14) gives s^2, and (3.15) results in zero. We also note that $P(0|n,c,s) = 0$ for $c = 2$ and $n < 2s$, and this is verified by (3.12) for the case $n = 2s - 1$. However, unlike Case A, we may have $P(0|n,c,s) > 0$ even for very small values of n if $c \geq 3$.

The alternative method for Case A also works here when $n \geq 2s - 1$. Again, we replace both $n+1-2s+k$ and $n-s+1$ in (3.2) by n. As before, we separate out the cases $s = k$ (for which the result is $n^{-(c-1)}$) and $c = 1$ (for which $K = s$ with probability one). For $0 < k < s$ and $c > 1$, we obtain

$$(3.16) \qquad P(k|n,c,s) = \frac{1}{n^{c-1}} \sum_{i=0}^{s-k-1} \binom{s-k-1}{i} (i+2)! \; S_c^{i+2} \quad \text{and}$$

$$(3.17) \qquad P(0|n,c,s) = 1 - \frac{1}{n^{c-1}} - \frac{1}{n^{c-1}} \sum_{k=1}^{s} \sum_{i=0}^{s-k-1} \binom{s-k-1}{i} (i+2)! \; S_c^{i+2} .$$

Note that by equating (3.9) to (3.2), (3.10) to (3.4), (3.16) to (3.11), and (3.17) to (3.12) we verify in each case the well-known finite difference identity $\Delta^k x^c]_{x=0} = k!\, S_c^k$ for the special case $k = 2$, which we used in (2.12).

For $n \leq 2s - 2$ and $c \geq 2$ the intersection need not be a set of successive points but can be the union of two or more such sets. In such cases (3.11), (3.16), (3.12), and (3.17) do not hold, and a single general formula has not been obtained. We have, however, obtained a method for finding these probabilities which is correct for a better (i.e., smaller) lower bound on n than $2s - 1$ (see above discussion), which we presently describe. We note first that (3.11) and (3.12) (or (3.16) and (3.17)) always hold for the trivial case $c = 1$.

Suppose the intersection consists of the union of two separated groups of successive points whose sizes are a and b, with $a \geq 1$ and $b \geq 1$. Then the probability distribution given by (3.11), (3.16), (3.12) and (3.17) incorrectly inserts the probability of this case into both $P(a|n,c,s)$ and $P(b|n,c,s)$ since this distribution deals only with contiguous intersections and assumes "overlap" is impossible. Since there are really $a+b$ points in the intersection, the old calculation must be corrected in the way outlined in the next paragraph.

Let $n = 2s-t$ for some $0 \leq t \leq s-1$. (The case $t = s$ also gives $n = s$ making the problem trivial, and we ignore it.) Then assuming a non-contiguous intersection, the maximum possible size of the intersection is t. Let $P_{a,b}$ represent the probability of a non-contiguous intersection

consisting of two groups whose sizes are a and b $(a \geq 1, \quad b \geq 1)$. Assuming we can find $P_{a,b}$ for all integer pairs (a,b) where $a \leq b$ and $a + b \leq t$, the corrected probabilities are given by the following procedure:

(3.18)

(i) Find $P(k|n,c,s)$ for $0 \leq k \leq s$ as given by (3.11) or (3.16) and (3.12) or (3.17).

(ii) For each (a,b) satisfying the above, subtract $P_{a,b}$ from both $P(a|n,c,s)$ and $P(b|n,c,s)$. (If $a = b$, the same subtraction is performed twice.)

(iii) For each (a,b) satisfying the above, add $P_{a,b}$ to $P(a+b|n,c,s)$.

(iv) For each (a,b) satisfying the above, add $P_{a,b}$ to $P(0|n,c,s)$ since the cases in question were added twice into numbers other than zero, and $P(0|n,c,s)$ was obtained by subtraction.

For $t = 0$ or $t = 1$, there are no pairs (a,b) for which $P_{a,b}$ must be computed. Hence the procedure (3.18) leaves (3.11) and (3.12) unchanged, as it must, because (3.11) and (3.12) are valid for $n \geq 2s - 1$. From the way (3.18) was carried out, it is clear that this procedure is correct whenever it is impossible for the intersection to consist of three or more separated groups of consecutive points. Hence (3.18) is always correct for $c = 2$. To ascertain how large n must be for $c \geq 3$ in order to assure that it will still be correct, we first consider two separated groups of points which are the intersection of two clusters. If a third cluster "splits up" one of these groups, the third cluster must start from one end of this group and go around the circle in the

opposite direction until it meets part of this same group of points at the other end. Since t is the maximum combined size of the original two groups (so that $t-1$ is the largest possible group) and since this third cluster must contain at least two points in some group plus all the points not in that group, there are at most $t-3$ points which are not in this third cluster. Hence $s \geq n - (t-3)$. Since $n = 2s-t$, we get $s \geq 2s-2t+3$ or $t \geq \frac{s+3}{2}$. But then $n \leq 2s - \frac{s+3}{2} = \frac{3s-3}{2}$, and these are the only conditions under which (3.18) fails. Since n is an integer, (3.18) holds whenever $n \geq [\frac{3s-3}{2}] + 1 = [\frac{3s-1}{2}]$, and certainly whenever $n \geq \frac{3s}{2}$, quite an improvement over the $2s - 1$ from (3.11).

To find $P_{a,b}$, we first define r by $a + b = t - r$ where $0 \leq r \leq t - 2$, which is possible by the conditions on a and b. If $r = 0$ then $n = 2s - t = (s-a) + (s-b)$. Thus any two points whose distance in one direction is $s - a$ have $s - b$ as their distance in the other direction. The number of ways two such points can be chosen is clearly $\frac{n}{2}(1 + \delta'_{ab})$, which is easily verified to be an integer, and the intersection of the two clusters whose right endpoints are some such pair of points consists of two separated groups whose sizes are a and b. Furthermore, adding any cluster distinct from these two will reduce the size of one of these groups. The probability that all of the c clusters chosen will be one of these two and that both of them will be chosen at least once is $(\frac{2}{n})^c - \frac{2}{n^c}$, which can be written as $Q_2(0)$ if we define for $j \leq r + 2$

$$(3.19)\qquad Q_j(r) = \begin{cases} \sum_{i=0}^{j} \binom{j}{i}(-1)^i \left(\frac{r+2-i}{n}\right)^c & \text{if } j \le c \\ 0 & \text{if } j > c\ . \end{cases}$$

$Q_j(r)$ represents the probability that all of the c selected clusters are among a specified set of $r+2$ clusters and that each cluster in a specified subset of size j out of these $r+2$ clusters is chosen at least once; it should be understood that $Q_j(r)$ depends on n and c despite the notation. Note that if $j = c$ the only freedom in selecting the clusters is their order so that $Q_c(r) = \frac{c!}{n^c}$ independently of r, and this is easily verified numerically for particular values of c. We remark, omitting the details, that by using finite difference calculus we can derive

$$(3.20)\qquad Q_j(r) = \frac{j!}{n^c} \sum_{\alpha=j}^{c} \binom{c}{\alpha}(r + 2 - j)^{c-\alpha} S_\alpha^j$$

where S_α^j are the Stirling numbers used in (2.12) and where the specification in (3.19) that $Q_j(r) = 0$ if $j > c$ is now superfluous. We now have for $r = 0$,

$$(3.21)\qquad P_{a,b} = \frac{n}{2}\,(1 + \delta'_{ab})\,Q_2(0)\ .$$

If $c = 2$, the number of points in a non-contiguous intersection must be exactly t; hence $r = 0$. For the cases $r > 0$, we can therefore assume $c \ge 3$. As in the case $r = 0$, we first select two points whose distance in one direction is $s-a$ and we use them as the right endpoints of clusters, yielding an intersection of a points on one side of the circle. We wish to introduce more clusters so as to reduce the size of

the intersection on the other side of the circle from $b + r$ to r. (It is presently $b + r$ since $(s - a) + \{s-(b+r)\} = 2s-a-b-r = n$ because $a+b+r = t$.) This can be done as follows:

(i) Starting from either of the points already selected, proceed around the circle in the direction in which the distance to the other point already selected is <u>not</u> $s-a$ until you have "traveled" a distance d_1 for some integer $0 \le d_1 \le r$, and select that point as a right cluster endpoint if that has not already been done (i.e., if $d_1 \neq 0$).

(ii) Start from the second of the original two points and repeat (i) using a distance $d_2 = r-d_1$.

In this manner we have selected either three distinct right cluster endpoints (if $d_1 = 0$ or $d_1 = r$) or four of them (if $1 \le d_1 \le r - 1$), and the total number of points in the intervals over which we "traveled" in (i) and (ii) above is $r + 2$. For $d_1 = 0$ or $d_1 = r$, the probability that all of the c clusters have one of these $r+2$ points as a right endpoint and that all of the three right cluster endpoints we have selected are included at least once is $Q_3(r)$. Similarly, for $1 \le d_1 \le r - 1$, the desired probability is $Q_4(r)$. Note that $Q_4(r) = 0$ if $c = 3$, which is necessary since all four required endpoints cannot be selected if the total number of clusters is only three. Similarly, $Q_4(r) = Q_3(r) = 0$ if $c = 2$. We also need the number of ways the original two points can be selected, which initially appears to be n. However, if $a = b$ every possible pair of groups of points of size $a = b$ in the intersection can

be obtained in two different ways by the above procedure since you can arrive at every such pair by beginning with the two endpoints that determine either member of the pair as one component of the intersection. So by slightly different methods from the case when $r = 0$, we obtain the same factor $\frac{n}{2}(1 + \delta'_{ab})$. For $r > 0$, we now have

$$(3.22) \qquad P_{a,b} = \frac{n}{2}(1 + \delta'_{ab})\{2Q_3(r) + (r-1)Q_4(r)\} .$$

Combining (3.21) and (3.22) yields (for any r)

$$(3.23) \qquad P_{a,b} = \frac{n}{2}(1 + \delta'_{ab})[\delta_{r0}\, Q_2(0) + \delta'_{r0}\, \{2Q_3(r) + (r-1)Q_4(r)\}]$$

where $\delta_{r0} = 1 - \delta'_{r0}$ is the Kronecker delta. The procedure (3.18) is now completely defined. (See insert following page.)

C. Linear Chain, Without Replacement.

For $c = 1$ we have $K = s$ with probability one, and for $s + 1 \le c \le n - s + 1$ we have $K = 0$ with probability one. Also, for any c we get $P(k|n,c,s) = 0$ if $k \ge s - c + 2$, which is equivalent to the condition $s - k = \alpha \le c - 2$. So we need only consider the cases $2 \le c \le s$ and $0 \le k \le s - c + 1$ (or, equivalently, $c - 1 \le \alpha \le s$). For $k > 0$, by the same reasoning as in Case A, we obtain for $n \ge 2s - k - 1$

$$(3.24) \qquad P(k|n,c,s) = \frac{n+1-2s+k}{\binom{n-s+1}{c}}\left\{\binom{s-k+1}{c} - 2\binom{s-k}{c} + \binom{s-k-1}{c}\right\}$$

with no need to include δ'_{sk} because we have assumed $k \le s - c + 1$ and $c \ge 2$ so that $k < s$. The result is zero as before if $n \le 2s - k - 1$,

Insert at end of 3.B.

For a numerical example when $n \geq 2s - 1$ (i.e., when (3.11) through (3.17) completely solve the problem), apply the principles of Section 4 to the example from Section 2, Case B. To illustrate the corrective procedure (3.18) when $n < 2s - 1$, we consider the case $n = 7$, $c = 4$, $s = 5$. Then using either (3.11) and (3.12) or (3.16) and (3.17) we first obtain, prior to the corrections specified by (3.18), the obviously incorrect results $P(0|7,4,5) = -\frac{26}{343}$, $P(1|7,4,5) = \frac{194}{343}$, $P(2|7,4,5) = \frac{110}{343}$, $P(3|7,4,5) = \frac{50}{343}$, $P(4|7,4,5) = \frac{14}{343}$, and $P(5|7,4,5) = \frac{1}{343}$. Note that these "probabilities" sum to one, which must be true, since (3.18) does not change this sum. In this example $t = 3$, so we use (3.23) to find only $P_{1,1}$ and $P_{1,2}$, whose values are $\frac{36}{343}$ and $\frac{14}{343}$, respectively. The adjustments called for by (3.18) now give the final result $P(0|7,4,5) = \frac{24}{343}$, $P(1|7,4,5) = \frac{108}{343}$, $P(2|7,4,5) = \frac{132}{343}$, $P(3|7,4,5) = \frac{64}{343}$, and the other two probabilities are unchanged. We now find $E(K) = \frac{625}{343}$ and $Var(K) = 110{,}498/117{,}649$. We note that for $n < 2s - 1$, the expectation as well as the variance is a quantity not obtainable by the methods of [2].

and this agrees with (3.24) for $n = 2s - k - 1$. Unlike Case A, we find that (3.24) is <u>always</u> valid for $n = 2s - k - 1$ since the case $k = s$ is not under consideration. Thus (3.24) is valid for every $1 \le k \le s - c + 1$ if $n \ge 2s - 2$. In <u>all</u> cases under consideration, (3.24) reduces to

$$(3.25) \qquad P(k|n,c,s) = \frac{n+1-s-\alpha}{\binom{n-s+1}{c}} \binom{\alpha-1}{c-2},$$

where $\alpha = s - k$, which helps verify the original contention that $\alpha \ge c - 1$ is necessary for a non-zero result. To sum these probabilities we use the well-known identity

$$(3.26) \qquad \sum_{\beta=a}^{b} \binom{\beta}{a} = \binom{b+1}{a+1} .$$

Summing on k in (3.25) from 1 to $s - c + 1$ (or, equivalently, on α from $c - 1$ to $s - 1$) and subtracting from one, we obtain for $n \ge 2s - 2$

$$(3.27) \qquad P(0|n,c,s) = 1 - \{c - \frac{s(c-1)}{n-s+1}\} \frac{\binom{s-1}{c-1}}{\binom{n-s}{c-1}} ;$$

the result is zero for $n < 2s$, and (3.27) bears this out for $n = 2s - 1$ and $n = 2s - 2$.

By defining either $S = \min(s-1,n-s)$ or $S = \min(s-1,n-s+1)$ for the same reasons that we made this definition prior to (3.5), we can compute $E(K|n,c,s)$ and $E(K^2|n,c,s)$ by a similar method to the one used in Case A. As before, both definitions of S will be valid for

the desired computations and will yield identical formulas. In either case, using (3.25) and (3.26) we obtain for $2 \le c \le s$ and $c - 1 \le \alpha \le s$,

$$(3.28)\quad E(K|n,c,s) = \sum_{k=\text{either of } \begin{cases} \max(1,2s-n) \\ \max(1,2s-n-1) \end{cases}}^{s-c+1} kP(k|n,c,s)$$

$$= \frac{1}{\binom{n-s+1}{c}} \sum_{\alpha=c-1}^{S} (s-\alpha)(n+1-s-\alpha)\binom{\alpha-1}{c-2}$$

$$= \frac{1}{\binom{n-s+1}{c}} \{s(n-s+1)\binom{S}{c-1} - (n+2)(c-1)\binom{S+1}{c} + c(c-1)\binom{S+2}{c+1}\}$$

$$= \begin{cases} \dfrac{(n+2)\binom{s}{c} - 2c\binom{s+1}{c+1}}{\binom{n-s+1}{c}} & \text{if } n \ge 2s-2 \\ \\ s(c+1) - c(n+2) + \dfrac{c^2+1}{c+1}(n-s+2) & \text{if } s \le n \le 2s-1 . \end{cases}$$

Also, by a similar computation,

$$(3.29)\quad E(K^2|n,c,s) = \frac{1}{\binom{n-s+1}{c}} \sum_{\alpha=c-1}^{S} (s-\alpha)^2(n+1-s-\alpha)\binom{\alpha-1}{c-2}$$

$$= \frac{1}{\binom{n-s+1}{c}} \{s^2(n+1-s)\binom{S}{c-1} - (c-1)[n+2+s(2n-s+3)]\binom{S+1}{c}$$

$$+ c(c-1)(n+s+4)\binom{S+2}{c+1} - (c-1)c(c+1)\binom{S+3}{c+2}\}$$

$$= \begin{cases} \dfrac{1}{\binom{n-s+1}{c}} [-(n+2)\binom{s}{c} + 2(n+c+3)\binom{s+1}{c+1} - 2(2c+1)\binom{s+2}{c+2}] \\ \qquad\qquad \text{if } n \ge 2s-2 \\ \\ s^2c - (c-1)(n+2+3s-s^2+2sn) + \dfrac{c(c-1)(n-s+2)}{(c+1)(c+2)}(2cs+c+n+3s+5) \\ \qquad\qquad \text{if } s \le n \le 2s-1. \end{cases}$$

Again, and for the same reasons as in Case A, the two final versions in both (3.28) and (3.29) agree when $n = 2s - 1$ or $n = 2s - 2$. Of course, we can now obtain

$$(3.30) \qquad \mathrm{Var}(K|n,c,s) = E(K^2|n,c,s) - E^2(K|n,c,s) .$$

It is easily verified that (3.28), (3.29), and (3.30) give the correct answers for $c = 1$, in which case $P(K = s) = 1$. (See insert following page.)

D. Circular Chain, Without Replacement.

As in Case B, we need to assume at first that n is large enough to avoid the possibility of having an intersection which is not a set of contiguous points. However, the number c of clusters chosen now has an effect on this possibility so that (if $c \geq 2$) we need only assume $n \geq 2s - c + 1$, which is not as strong as the old condition $n \geq 2s - 1$. The special cases $c = 1$, $s + 1 \leq c \leq n$, and $k \geq s - c + 2$ (or $\alpha \leq c - 2$) have the same answers as in Case C, and we again consider only $2 \leq c \leq s$ and $0 \leq k \leq s - c + 1$ (or $c - 1 \leq \alpha \leq s$). We replace both $n - s + 1$ and $n + 1 - 2s + k$ by n (to account for circularity) in (3.24) to obtain its proper analogue (for $k > 0$):

$$(3.31) \qquad P(k|n,c,s) = \frac{n}{\binom{n}{c}} \left\{\binom{s-k+1}{c} - 2\binom{s-k}{c} + \binom{s-k-1}{c}\right\} ,$$

which reduces in all cases under consideration to the analogue of (3.25),

$$(3.32) \qquad P(k|n,c,s) = \frac{n}{\binom{n}{c}} \binom{\alpha-1}{c-2} = \frac{c}{\binom{n-1}{c-1}} \binom{\alpha-1}{c-2} ,$$

where $\alpha = s - k$, which again helps verify the need for $\alpha \geq c - 1$ if

Insert at end of 3.C.

No useful example can be obtained here by trying to apply Section 4 to the example given in Case C of Section 2 ($n = 10$, $c = 4$, $s = 3$) since we have $P(0|n,c,s) = 1$ whenever $c > s$. Hence we consider the case $n = 10$, $c = 3$, $s = 4$. We use (3.25) to obtain $P(1|10,3,4) = \frac{8}{35}$ and $P(2|10,3,4) = \frac{5}{35}$. Then (3.27) yields $P(0|10,3,4) = \frac{22}{35}$. Therefore $E(K) = \frac{18}{35}$ and $Var(K) = \frac{556}{1,225}$, both of which are consistent with (3.28) and (3.30).

the result is to be non-zero. Using (3.26) to sum (3.32) on k from 1 to s - c + 1 (or, equivalently, on α from c - 1 to s - 1) and subtracting from one, we obtain

$$(3.33) \qquad P(0|n,c,s) = 1 - c\,\frac{\binom{s-1}{c-1}}{\binom{n-1}{c-1}} = 1 - n\,\frac{\binom{s-1}{c-1}}{\binom{n}{c}} \quad .$$

Exactly as we noted for (3.13) in Case B, (3.33) correctly yields zero when $c = 2$ and $n = 2s - 1$; this zero result remains correct when $c = 2$ for any $n < 2s$. Moreover, it also carries over from Case B that we may have $P(0|n,c,s) > 0$ even for small values of n when $c \geq 3$. Using (3.26) we now obtain for $n \geq 2s - c + 1$ and $2 \leq c \leq s$

$$(3.34) \qquad E(K|n,c,s) = \frac{c}{\binom{n-1}{c-1}} \sum_{\alpha=c-1}^{s-1} (s-\alpha)\binom{\alpha-1}{c-2} = c\,\frac{\binom{s}{c}}{\binom{n-1}{c-1}} = n\,\frac{\binom{s}{c}}{\binom{n}{c}} = s\,\frac{\binom{s-1}{c-1}}{\binom{n-1}{c-1}} ,$$

$$(3.35) \qquad E(K^2|n,c,s) = \frac{c}{\binom{n-1}{c-1}} \sum_{\alpha=c-1}^{s-1} (s-\alpha)^2\binom{\alpha-1}{c-2} = \frac{c}{\binom{n-1}{c-1}} \left[2\binom{s+1}{c+1} - \binom{s}{c}\right]$$

$$= \frac{n}{\binom{n}{c}} \left[2\binom{s+1}{c+1} - \binom{s}{c}\right], \quad \text{and}$$

$$(3.36) \qquad Var(K|n,c,s) = E(K^2|n,c,s) - E^2(K|n,c,s) = \frac{\binom{n-1}{c-1}\left\{2c\binom{s+1}{c+1} - c\binom{s}{c}\right\} - c^2\binom{s}{c}^2}{\binom{n-1}{c-1}^2}$$

where, as usual, we get s in (3.34), s^2 in (3.35), and zero in (3.36) when $c = 1$. Note that (3.36) has many equally simple alternate forms

obtainable by using the different forms of (3.34) and (3.35). We again emphasize that (3.31) through (3.36) are valid only for $n \geq 2s - c + 1$.

If $n \leq 2s - c$ and $c \geq 2$ we try to find the probability $P_{a,b}$ that the intersection consists of two separated groups of consecutive points whose sizes are a and b. Let $n = 2s - t$ where $0 \leq t \leq s - 1$. (We don't consider $t = s$ for the same reason as in Case B.) It is easily found that it is impossible for the intersection to consist of three or more separated groups of consecutive points unless $c \geq 3$ and $s \geq n - t + c$, which is the same lower bound on s as in Case B when $c = 3$ but which increases with c because the size of the maximum possible intersection of this type decreases as c increases. Hence a procedure analogous to (3.18) will always be valid for $c = 2$. For $c \geq 3$, since $n = 2s - t$, an intersection consisting of three or more separated groups of points can occur only when $s \geq 2s - 2t + c$ so that $t \geq \frac{s+c}{2}$. Then $n \leq 2s - \frac{s+c}{2} = \frac{3s-c}{2}$, and these are the only conditions under which our analogue to (3.18) will fail. Since n is an integer, these conditions cannot exist whenever $n \geq [\frac{3s-c}{2}] + 1 = [\frac{3s-c+2}{2}]$, and certainly (as before) whenever $n \geq \frac{3s}{2}$.

To find $P_{a,b}$ for any $a \geq 1$, $b \geq 1$, let $a + b = t - r$ where $0 \leq r \leq t - 2$. The maximum possible value of $a + b$ is t when $c = 2$, but this decreases by one every time we choose a new cluster. Hence $a + b \leq t - c + 2$ so that $P_{a,b} = 0$ if $r \leq c - 3$ and we can take $c - 2 \leq r \leq t - 2$. It also follows that if $t \leq c - 1$ then $a + b \leq t - c + 2 \leq 1$ and no possible pairs (a,b) exist. But then (3.31) through (3.36) apply and the problem is solved, so we can restrict ourselves to $c \leq t \leq s - 1$. In the remaining cases, the entire argument from Case B is valid here except that all quantities of the form y^c must

must be replaced by $\binom{y}{c}$. We now have

$$(3.37) \qquad Q_j(r) = \begin{cases} \sum_{i=0}^{j} \binom{j}{i} (-1)^i \dfrac{\binom{r+2-i}{c}}{\binom{n}{c}} & \text{if } j \le c \\ 0 & \text{if } j > c \ . \end{cases}$$

We can also define $Q_j(r) = 0$ if $r \le c - 3$ if we do not wish to treat this as a special case.

Given this new definition for $Q_j(r)$, (3.23) applies exactly as written, and (3.18) also applies with these new definitions of $Q_j(r)$ and $P_{a,b}$. We have now reduced the lower bound on n slightly further than we were able to in Case B. We remark only that $c = 2 \Leftrightarrow r = 0 \Leftrightarrow a + b = t$ (in Case B this was true only in the left-to-right direction as it was possible to have $r = 0$ when $c > 2$), and hence when we apply (3.23) to this case δ_{r0} and δ'_{r0} are respectively equivalent to δ_{c2} and δ'_{c2}. One final interesting property peculiar to the circular cases (B and D) is the possibility in these cases (when n is small) to have an empty intersection even though no two clusters are disjoint.

To first give an example for which (2.31) through (2.36) hold (i.e., for which $n \ge 2s - c + 1$), we again consider the case $n = 10$, $c = 3$, $s = 4$, which was introduced in Case C. From (3.32), we obtain $P(1|10,3,4) = \frac{2}{12}$ and $P(2|10,3,4) = \frac{1}{12}$; the result $P(0|10,3,4) = \frac{9}{12}$ is obtained from (3.33). Hence $E(K) = \frac{1}{3}$ and $Var(K) = \frac{7}{18}$, which are consistent with (3.34) and (3.36).

We now illustrate the corrective procedure dealing with the case $n \le 2s - c$ by considering $n = 7$, $c = 3$, $s = 5$. By using (3.32) and (3.33) we first obtain, prior to the corrections, the incorrect

results $P(0|7,3,5) = -\frac{7}{35}$, $P(1|7,3,5) = \frac{21}{35}$, $P(2|7,3,5) = \frac{14}{35}$, and $P(3|7,3,5) = \frac{7}{35}$. From the paragraph preceding (3.37), we see that we only need to find $P_{1,1}$. Using (3.37) and (3.23), we obtain $P_{1,1} = \frac{7}{35}$; the corrected probabilities are therefore $P(0|7,3,5) = 0$, $P(1|7,3,5) = \frac{7}{35}$, $P(2|7,3,5) = \frac{21}{35}$, and $P(3|7,3,5) = \frac{7}{35}$. Hence $E(K) = 2$ and $Var(K) = \frac{14}{35}$.

Again, the pattern is not unexpected, and we conjecture that it holds in general. That is, it seems that the circular case always has a smaller mean and variance than the corresponding linear case but that both the mean and the variance are asymptotically equivalent as $n \to \infty$.

4. RELATIONSHIP BETWEEN THE UNION AND THE INTERSECTION.

In any linear chain or in a circular chain large enough that the intersection must always be a set of consecutive points (i.e., $n \geq 2s - 1$ when working with replacement or $n \geq 2s - c + 1$ when working without replacement or $c = 1$ in either case), we assume that the intersection is of size $k > 0$. Then the two extreme clusters that include these k points have been chosen, and it is clear that the union of these two clusters, which consists of $2s - k$ points, is also the union of all the chosen clusters. Hence, if we denote the probability function associated with the union by P and the probability function associated with the intersection by P^*, we obtain for $k = 1,2,\ldots,s$,

$$(4.1) \qquad P^*(k|n,c,s) = P(2s-k|n,c,s) \ .$$

For $k = s$ (4.1) is trivially true even for the circular cases not presently under discussion. To compare (4.1) with our previous work for $1 \leq k \leq s - 1$, we find $P(2s-k|n,c,s)$ using (2.13). Noting that $r = 1$ is the only term in the sum (so that $k_1 = 2s - k$) and that $W_{2s-k} = n-2s + k + 1$, we obtain for $1 \leq k \leq s - 1$

$$(4.2) \qquad P(2s-k|n,c,s) = \frac{n-2s+k+1}{(n-s+1)^c} \sum_{j=2}^{\min(c,s-k+1)} A_0(s-k,\ j-1,s)j!\ S_c^j$$

$$= \frac{n-2s+k+1}{(n-s+1)^c} \sum_{j=2}^{\min(c,s-k+1)} \binom{s-k-1}{j-2}\ j!\ S_c^j \ ,$$

which corresponds to $P^*(k|n,c,s)$ as given by (3.9). Similar correspondences can be obtained easily between (2.7) and (3.2) and between (2.18) and (3.16), but are much harder to arrive at for (2.27) and (3.24) and for (2.29) and (3.30).

We can also use (4.1) to write an equally interesting result about the complementary probabilities:

$$(4.3) \qquad P^*(0|n,c,s) = \sum_{k=2s}^{n} P(k|n,c,s) \ .$$

For the special case $n = 2s$ there is therefore a 1-1 correspondence between the size of the union and that of the intersection. We illustrate (4.3) numerically for the linear chain with replacement using $n = 11$, $s = 3$, $c = 3$. In this case we use either (2.7) or (2.13) to obtain $P(6|11,3,3) = \frac{198}{729}$, $P(7|11,3,3) = \frac{210}{729}$, $P(8|11,3,3) = \frac{120}{729}$, $P(9|11,3,3) = \frac{60}{729}$, and $P(10|11,3,3) = P(11|11,3,3) = 0$ so that $P\{K \geq 6|11,3,3\} = \frac{588}{729}$. From (3.4), $P^*(0|n,c,s) = \frac{588}{729}$ as desired.

No general correspondences analogous to (4.1) or (4.3) exist in the circular cases in which the intersection need not be a set of contiguous points. For these cases we can, however, write

$$(4.4) \qquad P^*(0|n,c,s) \leq P(n|n,c,s)$$

since an empty intersection guarantees complete coverage but not vice-versa.

Again exluding the cases considered by the preceding paragraph, we observe that (denoting the union by K and the intersection by K^*) by combining (4.1) and (4.3) we obtain an inequality on the respective expectations:

(4.5) $E(K) \geq 2s - E(K^*)$

with equality if $n = 2s$. In fact, $P(K \geq 2s - K^*) = 1$ and if $n = 2s$ the corresponding equality has probability one. Thus if $n = 2s$, K and K^* also have equal variances.

5. RELATIONSHIP TO STIRLING NUMBERS OF THE SECOND KIND.

We rewrite (2.2) in summation notation as

$$(5.1) \qquad \Psi(n-s,c-1,s) = c! \sum_{t=1}^{u} A_0(n-s,c-t,s)\, P_{t-1}(c)$$

where u, the upper limit on this sum, need not be infinite as the notation in (2.2) might lead us to believe; we proceed to find u_0, the smallest (finite) u-value for which (5.1) already holds. It follows from the definition of the A_0-function that for non-negative x, y, and z, $A_0(x,y,z) = 0$ if and only if either $yz < x$ or $y > x$. Thus we can first conclude that u_0 satisfies the inequalities (i) $s(c-u_0) \geq n - s$ and (ii) $s(c-u_0-1) < n - s$. Simple algebra reduces this to $c - \frac{n}{s} < u_0 \leq c + 1 - \frac{n}{s}$, which is equivalent to $u_0 = [c + 1 - \frac{n}{s}] = c + 1 - [[\frac{n}{s}]]$ since u_0 must be an integer. Secondly, all terms in (5.1) are zero unless $c - t \leq n - s$ (i.e., unless $t \geq c - n + s$), so that the lower limit for t can be taken to be $\max(1, c - n + s)$. Using this fact, substituting u_0 for u, and applying the linear transformation $j = c - t + 1$, we obtain from (5.1) the equation

$$(5.2) \qquad \Psi(n-s,c-1,s) = c! \sum_{j=[[\frac{n}{s}]]}^{\min(c,n-s+1)} A_0(n-s,j-1,s)\, P_{c-j}(c) \ .$$

We now note that the right side of (2.14) without the outside factor $\frac{1}{(n-s+1)^c}$ represents the same quantity as (2.2), and hence also as (5.2). Hence, by equating these quantities, we obtain

$$(5.3) \qquad c! \sum_{j=[[\frac{n}{s}]]}^{\min(c,n-s+1)} A_0(n-s,j-1,s) P_{c-j}(c) = \sum_{j=[[\frac{n}{s}]]}^{\min(c,n-s+1)} A_0(n-s,j-1,s)\, j!\, S_c^j$$

for all values of n, c, and s. In particular, fix c and any integer $1 \le m \le c$, and consider the case $n = m$ and $s = 1$. Then (5.3) yields

$$(5.4) \qquad \sum_{j=[[m]]}^{\min(c,m)} A_0(m-1,j-1,1)c!\, P_{c-j}(c) = \sum_{j=[[m]]}^{\min(c,m)} A_0(m-1,j-1,1)\, j!\, S_c^j .$$

But $m \le c$ by hypothesis, so $\min(c,m)=m$ and both sides of (5.4) sum only a single term, namely $j = m$. Also, $A_0(m-1,m-1,1) = 1$ for any m; hence (5.4) reduces to $c!P_{c-m}(c) = m!\, S_c^m$ (or $S_c^m = \frac{c!}{m!} P_{c-m}(c)$) for any $1 \le m \le c$, so that our polynomials $P_i(c)$ for $i = 0, 1, 2, \ldots$ generate the Stirling numbers of the second kind, which are always integers. For example,

$$P_{7-3}(c) = P_4(c) = \left\{\frac{\binom{c-4}{1}}{5!} + \frac{2\binom{c-4}{2}}{2!4!} + \frac{\binom{c-4}{2}}{(3!)^2} + \frac{3\binom{c-4}{3}}{3!(2!)^2} + \frac{\binom{c-4}{4}}{(2!)^4}\right\}$$

$$= \frac{1}{5760}(15c^4 - 210c^3 + 1085c^2 - 2442c + 2008); \text{ thus } P_4(7) = 2064/5760 = 43/120$$

and $\frac{7!}{3!} P_4(7) = 840\left(\frac{43}{120}\right) = 7(43) = 301$, which is the Stirling number S_7^3.

We believe that these polynomials represent a new way of generating these Stirling numbers. However, even if it turns out otherwise, we believe that the polynomials themselves form a sequence (with P_i having degree i) which is of interest per se in view of the relationship obtained in this section.

Acknowledgment

The authors wish to thank Claudia Tysdel for her painstaking efforts in the typing of this paper.

for all values of n, c, and s. In particular, fix c and any integer $1 \le m \le c$, and consider the case $n = m$ and $s = 1$. Then (5.3) yields

$$(5.4) \qquad \sum_{j=[[m]]}^{\min(c,m)} A_0(m-1,j-1,1)c!\, P_{c-j}(c) = \sum_{j=[[m]]}^{\min(c,m)} A_0(m-1,j-1,1)\, j!\, S_c^j .$$

But $m \le c$ by hypothesis, so $\min(c,m)=m$ and both sides of (5.4) sum only a single term, namely $j = m$. Also, $A_0(m-1,m-1,1) = 1$ for any m; hence (5.4) reduces to $c!P_{c-m}(c) = m!\, S_c^m$ (or $S_c^m = \frac{c!}{m!} P_{c-m}(c)$) for any $1 \le m \le c$, so that our polynomials $P_i(c)$ for $i = 0, 1, 2, \ldots$ generate the Stirling numbers of the second kind, which are always integers. For example,

$$P_{7-3}(c) = P_4(c) = \left\{\frac{\binom{c-4}{1}}{5!} + \frac{2\binom{c-4}{2}}{2!4!} + \frac{\binom{c-4}{2}}{(3!)^2} + \frac{3\binom{c-4}{3}}{3!(2!)^2} + \frac{\binom{c-4}{4}}{(2!)^4}\right\}$$

$$= \frac{1}{5760}(15c^4 - 210c^3 + 1085c^2 - 2442c + 2008); \text{ thus } P_4(7) = 2064/5760 = 43/120$$

and $\frac{7!}{3!} P_4(7) = 840(\frac{43}{120}) = 7(43) = 301$, which is the Stirling number S_7^3.

We believe that these polynomials represent a new way of generating these Stirling numbers. However, even if it turns out otherwise, we believe that the polynomials themselves form a sequence (with P_i having degree i) which is of interest per se in view of the relationship obtained in this section.

Acknowledgment

The authors wish to thank Claudia Tysdel for her painstaking efforts in the typing of this paper.

References

[1] David, F. N. and Barton, D. E. (1962), Combinatorial Chance, pp. 227-30, Charles Griffin Co., Ltd., London.

[2] Hoel, D. G., Sobel, M., and Uppuluri, V. R. R. (1973) "Cluster Problems in One Dimension", Submitted to Jour. of Applied Probability.

[3] Olkin, I. and Sobel, M. (1965), "Integral Expressions for Tail Probabilities of the Multinomial and Negative Multinomial Distributions", Biometrika, 52: 167 - 79.